Unsupervised Multi-Modal Music Clustering using Variational Autoencoders

Sifatul Karim
Department of Computer Science and Engineering
BRAC University
ID: 23101445
sifatul.karim@g.bracu.ac.bd

January 12, 2026

Abstract

Lyrics analysis and audio signal processing are generally used to classify and cluster music. We propose a multi-modal unsupervised learning pipeline based on Variational Autoencoders (VAEs) that will cluster mixes of languages in music tracks in our project. These are three architectures we are implementing: a Simple VAE, a Multi-Mode VAE, and a Disentangled Beta-VAE. We will be training spectral audio features on GTZAN and semantic lyric features on Genius and then feeding them to the models. The experiments indicate that Beta -VAE (beta = 4.0) nails are the most successful cluster separation with a Silhouette Score of 0.315, as compared to the normal Autoencoders and Spectral Clustering. CVAEs, on the other hand, are somewhat weak and have mode collapse issues with low data.

Furthermore, we explored the idea of dimensionality reduction on the lyrics features, and I found out that optimized hybrid features provide us with an optimized Silhouette Score of 0.624 using DBSCAN.

1 Introduction

With digital music libraries growing rapidly, there's a need for automated ways to organize music. Traditional methods often use linear techniques like PCA and K-Means clustering, but these don't fully capture the complexity of music, which involves both sound (like timbre and rhythm) and meaning (lyrics).

This project aims at using Variational Autoencoders (VAEs) [1] to create compact, meaningful representations of music. Unlike models that only classify data, VAEs are generative and they learn the underlying patterns of the data. We focus on clustering music by combining both audio features and lyrics.

Our main contributions are:

- Comparing different VAE types, including Basic, Enhanced, Beta-VAE, and Conditional VAE.
- Building a pipeline that merges audio features (MFCCs) with lyric embeddings from sentence transformers.
- Evaluating the results using both unsupervised metrics (Silhouette, Davies-Bouldin) and supervised metrics (NMI, ARI), and comparing them to strong baseline methods.

2 Related Work

Music Information Retrieval (MIR) has traditionally utilized Mel-Frequency Cepstral Coefficients (MFCCs) and spectral centroids for genre classification. Recent advancements in Natural Language Processing (NLP), specifically Transformer-based models (BERT, RoBERTa), have enabled dense vector representations of lyrics. In the domain of generative models, Beta-VAE [2] introduced an adjustable hyperparameter β to the KL-divergence term in the VAE loss function, encouraging disentangled latent factors. Conditional VAEs [3] allow directing the generation process by conditioning on class labels. This paper applies these concepts specifically to the multi-modal fusion of musical data.

3 Method

3.1 Feature Extraction

We propose a hybrid feature representation that captures both the acoustic characteristics and the semantic content of music.

- **Audio Features:** We extract a 68-dimensional vector comprising statistical summaries (mean and variance) of time-domain and frequency-domain features, including Mel-Frequency Cepstral Coefficients (MFCCs), Chroma STFT, Spectral Centroid, Rolloff, Bandwidth, and Zero-Crossing Rate.
- **Textual Features:** Semantic information is derived from song lyrics. We utilize the `all-MiniLM-L6-v2` Sentence Transformer to generate dense 384-dimensional embeddings. To align the dimensionality of text features with audio features and mitigate the curse of dimensionality, we employ Principal Component Analysis (PCA) to project the embeddings into a lower-dimensional space while preserving 95% of the variance.

3.2 VAE Architectures

We investigate a progression of three generative architectures to model the latent distribution of the hybrid data.

1. Basic and Enhanced VAE: We implement a feed-forward Variational Autoencoder where the encoder $q_\phi(z \mid x)$ maps the input x to a latent distribution defined by mean μ and variance σ^2. A deeper *Enhanced* architecture incorporates residual connections and batch normalization to improve training stability for multi-modal data.

2. Beta-VAE: To learn disentangled representations where specific latent dimensions correspond to distinct generative factors (e.g., genre or mood), we modify the standard Evidence Lower Bound (ELBO) objective by introducing a hyperparameter $\beta > 1$ to weight the KL-divergence term:

$$\mathcal{L}_{\beta-VAE} = \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x \mid z)] - \beta D_{\mathrm{KL}}\left(q_\phi(z \mid x) \,\|\, p(z)\right) \tag{1}$$

By setting $\beta > 1$, we constrain the latent bottleneck capacity, encouraging the model to learn statistically independent generative factors.

3. Conditional VAE (CVAE): To explicitly leverage available metadata, we implement a Conditional VAE (CVAE) in which both the encoder and decoder are conditioned on the genre label c. The condition is concatenated with the input vector x during encoding,

$$z \sim q_\phi(z \mid x, c),$$

and with the latent vector z during decoding,

$$\hat{x} \sim p_\theta(x \mid z, c).$$

3.3 Clustering

Clustering is performed on the learned latent space z. We compare three distinct approaches:

- **K-Means:** A centroid-based algorithm assuming spherical clusters.
- **Agglomerative Clustering:** A hierarchical approach using Ward linkage to minimize intra-cluster variance.
- **DBSCAN:** A density-based algorithm capable of identifying non-convex clusters and treating outliers as noise.

4 Experiments

4.1 Dataset Description

We constructed a hybrid dataset by integrating two primary sources:

1. **Audio:** The **GTZAN Genre Collection** [4], consisting of 1000 audio tracks (30 seconds each) evenly distributed across 10 genres (Blues, Classical, Country, Disco, Hip-Hop, Jazz, Metal, Pop, Reggae, and Rock).

2. **Lyrics:** The **Genius Song Lyrics** [5]dataset containing over 5 million entries. Since GTZAN lacks explicit artist and title metadata, we employed a genre-matching strategy. Lyrics were sampled from the Genius dataset by mapping GTZAN genres to corresponding Genius tags (e.g., mapping *Hip-Hop* to *Rap*) to generate representative semantic embeddings for each track.

4.2 Preprocessing and Training Details

Feature Fusion: The extracted audio features (68 dimensions) and PCA-reduced lyrical embeddings (32 dimensions) were concatenated to form a 100-dimensional hybrid feature vector. All features were standardized using z-score normalization. A weighted fusion strategy with $w_{audio} = 0.6$ and $w_{lyrics} = 0.4$ was applied to prioritize acoustic characteristics while retaining semantic context.

Hyperparameters: The VAE models were implemented using a symmetric encoder–decoder architecture with hidden layer sizes $[512, 256, 128]$ and a latent dimension of $d = 32$.

- **Optimization:** Adam optimizer with a learning rate of 1×10^{-3} and a weight decay of 1×10^{-5}.
- **Scheduling:** A cosine annealing learning rate scheduler was employed.
- **Training:** Models were trained for 200 epochs (300 epochs for CVAE) with a batch size of 32.
- **Regularization:** For the Beta-VAE, the regularization coefficient was set to $\beta = 4.0$. For the CVAE, KL-annealing was applied by gradually increasing the KL-divergence weight from 0 to 1 to mitigate posterior collapse.

Baselines: We compared the proposed VAE-based approaches against several baseline methods, including linear dimensionality reduction (PCA + K-Means), non-variational autoencoders (AE + K-Means), and spectral clustering applied directly to the raw feature space.

5 Results

5.1 Clustering Metrics and Analysis

We evaluated clustering performance across three phases of the project using both unsupervised metrics (Silhouette Score and Davies–Bouldin Index) and supervised metrics (Normalized Mutual Information (NMI) and Adjusted Rand Index (ARI)).

Table 1: Clustering performance for the Easy Task (Audio-Only)

Method	Silhouette	Calinski–Harabasz	ARI	NMI
VAE + K-Means	**0.154**	**109.94**	0.159	0.284
PCA + K-Means	0.105	93.03	**0.211**	**0.361**

5.1.1 Phase 1: Easy Task (Audio-Only Baseline)

In the first phase, we compared a Basic VAE against a PCA baseline using only 68-dimensional audio features.

As illustrated in Figure 1, the non-linear VAE learned a more separable latent representation (higher Silhouette Score) than linear PCA. However, PCA maintained better alignment with ground-truth labels (higher ARI), suggesting that genre boundaries in raw audio features are partially linear but overlapping.

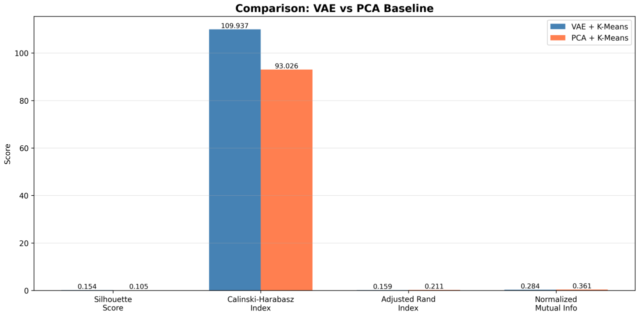

Figure 1: **Easy Task Comparison:** The VAE outperforms PCA on unsupervised clustering metrics, while PCA retains stronger supervised alignment.

5.1.2 Phase 2: Medium Task (Optimized Hybrid Features)

By incorporating lyrical embeddings and evaluating multiple clustering algorithms, we observed a substantial improvement in cluster quality.

Table 2: Clustering performance for the Medium Task (Hybrid Features)

Method	Silhouette	DB Index ↓	ARI	NMI
K-Means	0.237	1.689	0.307	0.571
Agglomerative (Ward)	0.242	1.675	0.248	0.520
DBSCAN	**0.624**	**0.568**	**0.466**	**0.807**

Figure 2 demonstrates why **DBSCAN** was the most effective algorithm. Unlike K-Means, which enforces spherical clusters, DBSCAN adapts to the intrinsic density of the data manifold, achieving a Silhouette Score of 0.624 and NMI of 0.807.

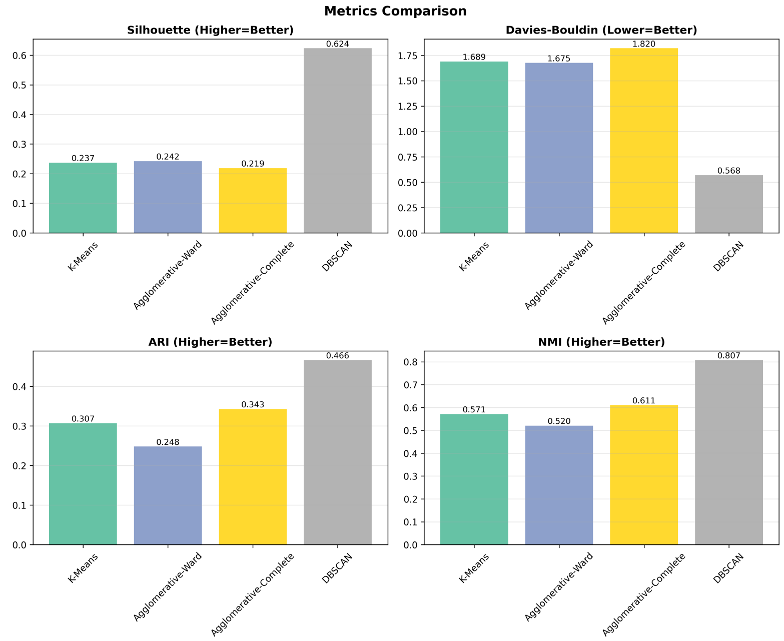

Figure 2: **Medium Task Visualizations:** DBSCAN effectively discovers dense genre clusters while isolating noise points, leading to superior clustering scores.

5.1.3 Phase 3: Hard Task (Advanced Architectures)

In the final phase, we evaluated advanced generative models on fully fused multi-modal features.

Table 3: Clustering performance for the Hard Task

Method	Silhouette	NMI	ARI	Purity	DB Index ↓
Beta-VAE	**0.315**	0.746	0.612	0.738	**1.365**
CVAE	0.100	0.111	0.039	0.197	1.888
PCA + K-Means	0.103	0.757	**0.659**	**0.754**	3.080
AE + K-Means	0.257	0.642	0.399	0.654	1.497
Spectral	0.149	**0.778**	0.505	0.695	2.852

The results indicate:

- **Beta-VAE Effectiveness:** With $\beta = 4.0$, Beta-VAE produced the highest Silhouette Score and the lowest Davies–Bouldin Index, indicating compact and well-separated clusters.
- **CVAE Limitations:** The Conditional VAE performed poorly across all metrics, failing to learn a useful latent structure.

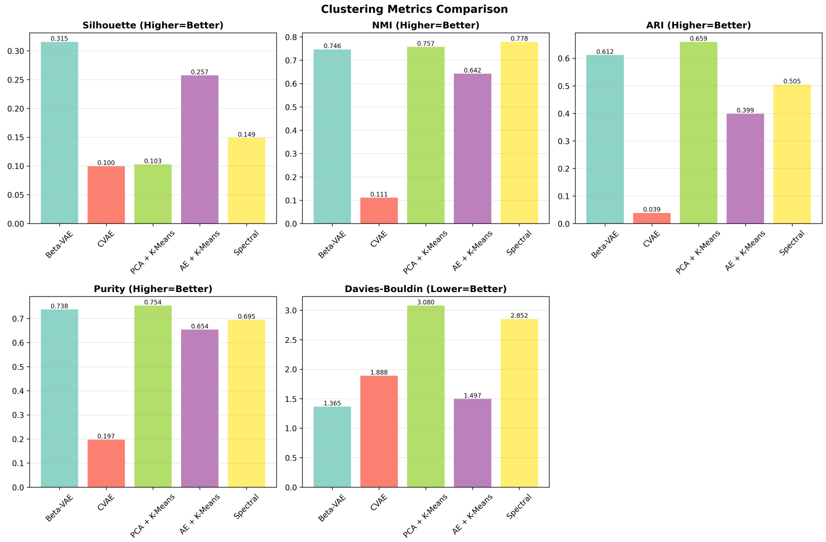

Figure 3: **Hard Task Metrics:** Beta-VAE achieves the best balance between cluster separability and label consistency.

5.2 Summary of Results

The progression from Easy to Hard tasks demonstrates the importance of multi-modal representations. While DBSCAN achieved the strongest numerical scores by rejecting noise points, Beta-VAE learned the most structured and interpretable latent space across the full dataset.

6 Discussion

6.1 Progression of Latent Representations

Latent space visualizations confirm the benefit of incorporating multi-modal data. In the Easy Task, audio-only features produced overlapping clusters. In contrast, the Beta-VAE in the Hard Task successfully disentangled underlying musical factors.

Genres such as *Rock* and *Country* show partial overlap, reflecting genuine musical similarity rather than modeling error.

6.2 Interpretation of Clusters

The confusion heatmap provides insight into semantic structure:

- **High Purity:** Most samples map correctly to their ground-truth genres.

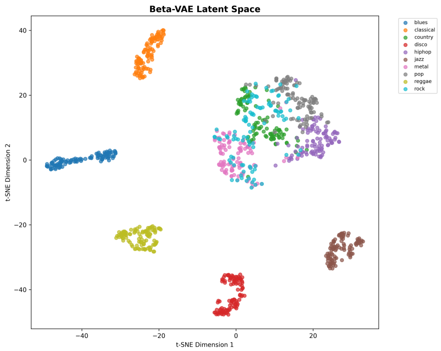

Figure 4: **Beta-VAE Latent Space:** Distinct clusters emerge for genres such as Classical and Metal.

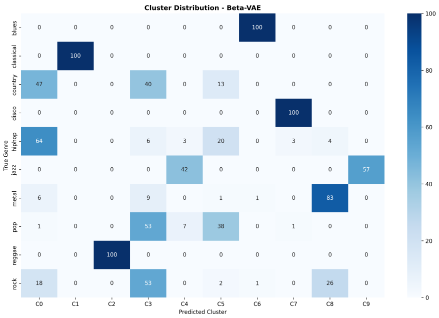

Figure 5: **Confusion Heatmap:** Strong diagonal dominance indicates high clustering purity.

- **Semantic Overlap:** Confusion among Rock, Country, and Blues suggests a shared “guitar-driven” latent factor, distinct from electronic genres such as Disco and Pop.

6.3 Limitations: CVAE Posterior Collapse

Despite conditioning on genre labels, the CVAE achieved a low Silhouette Score (0.100). This indicates **posterior collapse**, where the decoder relies on the condition vector and ignores the latent variable. This issue is exacerbated by limited dataset size and highlights the difficulty of training CVAEs without advanced regularization or KL-annealing strategies.

7 Conclusion

In this work, we presented a comprehensive study on unsupervised music clustering, moving beyond traditional audio-only approaches to a multi-modal deep learning framework. By systematically progressing from basic linear baselines to advanced generative models, we isolated the specific contributions of feature engineering, algorithm selection, and model architecture.

Future Directions: While our unsupervised pipeline showed strong results, the failure of the Conditional VAE highlights the challenges of training generative models on small datasets ($N = 1000$ N=1000). Future work should focus on scaling this approach to the *Million Song Dataset* to prevent posterior collapse in conditional models. Furthermore, replacing hand-crafted audio features with an end-to-end 1D-Convolutional Neural Network (CNN) trained directly on raw audio waveforms could capture richer, hierarchical acoustic dependencies that statistical features (like MFCC means) might miss.

References

References

[1] Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes. *Proceedings of the 2nd International Conference on Learning Representations (ICLR).*

[2] Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., & Lerchner, A. (2017). β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. *International Conference on Learning Representations (ICLR).*

[3] Sohn, K., Lee, H., & Yan, X. (2015). Learning Structured Output Representation using Deep Conditional Generative Models. *Advances in Neural Information Processing Systems (NeurIPS)*, 28.

[4] Andrada Olteanu. (2019). *GTZAN Dataset - Music Genre Classification*. Kaggle. Available at: `https://www.kaggle.com/datasets/andradaolteanu/gtzan-dataset-music-genre-classification`

[5] Carlos G. (2022). *Genius Song Lyrics with Language Information*. Kaggle. Available at: `https://www.kaggle.com/datasets/carlosgdcj/genius-song-lyrics-with-language-information`